ОЦЕНКА ВЛИЯНИЯ ЗАСОРЕНИЙ ОБУЧАЮЩЕЙ ВЫБОРКИ НА ТОЧНОСТЬ АЛГОРИТМОВ КРЕДИТНОГО СКОРИНГА

А. Е. Алексеев, В. С. Дежемесова

ВВЕДЕНИЕ

Целью кредитного скоринга является управление кредитным риском, а именно автоматизация принятия решений по выдаче банковских кредитов, а также условиям кредитования. Кредитный скоринг осуществляется с помощью компьютерных систем, обеспечивающих автоматическую классификацию потенциальных заемщиков коммерческого банка на основе доступной информации по степени кредитоспособности. Основным компонентом систем кредитного скоринга являются реализованные в них статистические модели и методы классификации неоднородных многомерных данных. К числу наиболее часто используемых в системах кредитного скоринга методов классификации относятся дискриминантный анализ (в предположении о совместном нормальном распределении анализируемых признаков) и методы классификации на основе логит-модели и пробит-модели бинарного и множественного выбора [1].

Целью данной статьи является сравнительный анализ точности классификации с помощью алгоритма дискриминантного анализа и алгоритма классификации на основе логит-модели бинарного выбора. Кроме того, исследуется влияние засорений в выборке на процент точных классификаций. Для тестирования используются реальные данные по заемщикам белорусского коммерческого банка.

1. МАТЕМАТИЧЕСКАЯ ПОСТАНОВКА ЗАДАЧИ КРЕДИТНОГО СКОРИНГА

Пусть потенциальный заемщик банка характеризуется некоторым набором показателей, из которых образован N-мерный вектор признаков $x = (x_1,...,x_N)^T \in \Re^N$. Предположим, что целью кредитного скоринга является классификация потенциальных заемщиков банка на L=2 класса: класс Ω_0 надежных заемщиков, относительно которых ожидается выполнение кредитных обязательств, и класс Ω_1 ненадежных заемщиков, относительно которых ожидается невыполнение кредитных обя-

зательств. В выдаче кредита заемщикам из класса Ω_1 может быть отказано. Данную задачу можно сформулировать в терминах теории статистических решений следующим образом.

Пусть в пространстве $\Re^N$ регистрируются случайные наблюдения $x = x(\omega) \in \Re^N$ над объектами $\omega \in \Omega$, принадлежащими к $L = 2$ классам $\{\Omega_0, \Omega_1\}$, удовлетворяющим условиям:

$$\Omega_0 \cap \Omega_1 = \varnothing,\ \Omega_0 \cup \Omega_1 = \Omega.$$

Истинный номер класса $d^0 = d^0(\omega) \in S = \{0,1\}$, к которому принадлежит наблюдение $x = x(\omega)$, является дискретной случайной величиной с распределением вероятностей:

$$P\{d^0(\omega) = i\} = \pi_i > 0,\ i \in S;\ \pi_0 + \pi_1 = 1, \qquad (1)$$

где $\pi_0, \pi_1 = 1 - \pi_0$ – априорные вероятности классов.

Случайный вектор признаков $x = x(\omega)$ для объектов из фиксированного класса $\Omega_i \in S$ $(d^0 = i)$ описывается некоторой условной плотностью распределения $p_i(x)$, а безусловная плотность распределения случайного вектора $x = x(\omega)$ (плотность смеси распределений) определяется выражением:

$$p(x) = \pi_0 p_0(x) + (1 - \pi_0) p_1(x). \qquad (2)$$

Задача кредитного скоринга заключается в отнесении заемщика коммерческого банка $\omega \in \Omega$ к одному из классов $\{\Omega_i\}_{i \in S}$ по совокупности его признаков $x = x(\omega)$, то есть в оценивании неизвестного (ненаблюдаемого) номера класса $d^0 = d^0(\omega) \in S$ для заемщика ω по известному значению его показателей $x = x(\omega) \in \Re^N$.

На практике вероятностные характеристики $\{\pi_i, p_i(x)\}_{i \in S}$ классов $\{\Omega_i\}_{i \in S}$ частично или полностью не известны. Однако банк может располагать кредитной базой данных, включающей информацию о заемщиках, для которых ранее выдавались кредиты и классификация которых на классы $\{\Omega_0, \Omega_1\}$ к текущему моменту точно известна. Обозначим через $X = \{x_1, ..., x_n\}$ выборку векторов признаков, соответствующих таким заемщиков коммерческого банка и будем называть ее классифицированной обучающей выборкой объема n. Будем предпола-

гать, что выборка $X = \{x_1,...,x_n\}$ образована из n независимых в совокупности случайных векторов $x_1,...,x_n$ с плотностью распределения $p(x)$ вида (2). Поскольку обучающая выборка является классифицированной, то для каждого наблюдения $x_t = x(\omega_t)$ из обучающей выборки $X = \{x_1,...,x_n\}$ точно известен номер класса $d_t^0 = d^0(\omega_t) \in S (t = 1, \ldots, n)$.

Обучающая выборка X используется на «этапе обучения» для статистического оценивания вероятностных характеристик $\{\pi_i, p_i(x)\}_{i \in S}$ и построения решающего правила классификации, которое затем применяется на «этапе экзамена» для классификации новых заемщиков по соответствующим наблюдениям $x_{n+1}, x_{n+2}, \ldots$.

Методы построения и конкретный вид решающего правила классификации зависят от дополнительных модельных предположений относительно вероятностной модели наблюдений, которые в свою очередь, обусловлены особенностями реально наблюдаемых показателей. Рассмотрим случай, когда вектор признаков $x = (x_1,...,x_N)^T \in \Re^N$ имеет нормальный закон распределения.

2. АЛГОРИТМ КРЕДИТНОГО СКОРИНГА НА ОСНОВЕ ДИСКРИМИНАНТНОГО АНАЛИЗА ГАУССОВСКИХ СЛУЧАЙНЫХ ВЕКТОРОВ

Пусть условные плотности $\{p_i(\cdot)\}_{i \in S}$, описывающие классы $\{\Omega_i\}_{i \in S}$ являются плотностями *N*-мерного нормального закона распределения и различаются для разных классов значениями параметров:

$$p_i(x) = n_N(x|\mu_i, \Sigma_i),\ x \in \Re^N,\ i \in S, \tag{3}$$

где для гауссовского случайного вектора наблюдений $x \in \Re^N$ из класса $\Omega_i\ (d^0 = i)$: $\mu_i = \mathbf{E}\{x|d^0 = i\}$ – вектор условного математического ожидания, $\Sigma_i = \mathbf{E}\{(x - \mu_i)(x - \mu_i)^{'}|d^0 = i\}$ – невырожденная условная ковариационная $(N \times N)$-матрица $(|\Sigma_i| \neq 0)$.

Сформулируем вначале оптимальное в смысле минимума вероятности ошибки решающее правило (известное как байесовское решающее правило) для двух классов $L = 2,\ S = \{0,1\}$.

Оптимальное в смысле минимума вероятности ошибки решение о принадлежности заемщика с характеристиками x к классу с номером $\hat{d} \in S = \{0,1\}$ выносится с помощью квадратичного байесовского решающего правила (БРП) вида [4]:

$$\hat{d} \equiv \hat{d}(x) = \arg\min_{i \in S}((x - \mu_i)' \Sigma_i^{-1}(x - \mu_i) + \ln|\Sigma_i| - 2\ln\pi_i). \quad (4)$$

В случае, когда ковариационные матрицы случайных векторов признаков для всех классов равны ($\Sigma_0 = \Sigma_1 = \Sigma, |\Sigma| \neq 0$) модель (1)-(4) часто называется моделью Фишера. Для модели Фишера БРП (4) принимает вид:

$$\hat{d}(x) = \arg\min_{i \in S}((x - \mu_i)' \Sigma^{-1}(x - \mu_i) - 2\ln\pi_i). \quad (5)$$

БРП (5) допускает эквивалентное представление в виде:

$$\hat{d}(x) = \mathbf{U}(G_0(x)) + 1, \; x \in \Re^N, \quad (6)$$

где $\mathbf{U}(w) = \begin{cases} 0, \text{ если } w < 0, \\ 1, \text{ если } w \geq 0 \end{cases}$ – единичная функция Хевисайда,

$$G_0(x) = \beta^T x + \beta_0,$$

$$\beta = (\mu_1 - \mu_0)^T \Sigma^{-1} \in \Re^N, \; \beta_0 = -\frac{1}{2}(\mu_1 + \mu_0)^T \Sigma^{-1}(\mu_1 - \mu_0) - \ln\frac{\pi_0}{1 - \pi_0}. \quad (7)$$

– линейная дискриминантная функция Фишера.

При равновероятных классах ($\pi_0 = \pi_1 = 1/2$) вероятность ошибки $\hat{P}$ для БРП (6), (7) вычисляется по формуле:

$$\hat{P} = \Phi\left(-\frac{\Delta}{2}\right), \; \Delta = \sqrt{\left(\mu_1 - \mu_0\right)^T \Sigma^{-1}\left(\mu_1 - \mu_0\right)}, \quad (8)$$

где Δ – межклассовое расстояние Махаланобиса, характеризующее степень разделимости классов, $\Phi(\cdot)$ – функция распределения стандартного нормального закона.

В случае, когда вероятностные характеристики $\{\pi_i, p_i(x)\}_{i \in S}$ классов $\{\Omega_i\}_{i \in S}$ неизвестны и имеется классифицированная обучающая выборка $X = \{x_1, ..., x_n\}$, используется подстановочное байесовское решающее правило (ПБРП), которое получается подстановкой в БРП (4) (или (5))

несмещенных статистических оценок неизвестных характеристик $\{\pi_i, \mu_i, \Sigma_i\}_{i \in S}$ $(i \in S)$.

Для исследования прогностической способности модели вычисляются ожидаемые значения номера класса $\hat{d}(x_t)$ и сравниваются с истинным номером $d^0(x_t)$, к которому принадлежит

t-й заемщик $(t = 1, \ldots, n)$. По результатам классификации выборки $X = \{x_1, ..., x_n\}$ находятся оценки вероятностей ошибок первого и второго рода ($\hat{P}_1$ и $\hat{P}_2$, соответственно):

$$\hat{P}_1 = P\{\hat{d}(x_t) \neq d^0(x_t) \mid d^0(x_t) = 1\},\ \hat{P}_2 = P\{\hat{d}(x_t) \neq d^0(x_t) \mid d^0(x_t) = 0\},$$

где $\hat{P}_1$ – вероятность ошибочного признания «проблемного» заемщика «непроблемным» (потери типа «прямых убытков»), $\hat{P}_2$ – вероятность ошибочного признания «непроблемного» заемщика «проблемным» (потери типа «упущенной выгоды»).

Очевидно, что чем меньше значения вероятностей ошибок, тем лучше прогностические способности модели.

3. ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ АЛГОРИТМОВ КРЕДИТНОГО СКОРИНГА

В этом пункте описанный выше скоринговый алгоритм применяется для классификации 40 заемщиков белорусского коммерческого банка.

Для анализа финансового состояния заемщика банка используются следующие балансовые коэффициенты: коэффициент абсолютной (мгновенной) ликвидности юридического лица (K2), коэффициент обеспеченности собственными средствами (K5), коэффициент финансовой независимости (автономии) (K7), коэффициент финансовой напряженности (K9).

Обучающая выборка включает 40 наблюдений. Таким образом, имеем $L = 2,\ N = 4,\ n = 40$.

Для классификации заемщиков банка, кроме описанного выше алгоритма дискриминантного анализа, используется классификация на основе логит-модели бинарного выбора [2], в которой вектор объясняющих переменных также сформирован из коэффициентов K2, K5, K7, K9. Вероятности ошибочных классификаций (ошибки первого и второго рода, а также безусловная вероятность ошибки в процентах) для различных алгоритмов для исходной выборки, а также при удалении аномальных наблюдений представлены в табл. ниже.

Таблица

Оценивание вероятности ошибочной классификации

Метод	До/после удаления аномальных наблюдений	$\hat{P}_1$	$\hat{P}_2$	$\hat{P}$
Алгоритм дискриминантного анализа	Исходная выборка	14.8	46.15	25
	Выборка без аномальных наблюдений	8.00	42.8	15.6
Алгоритм логит-модели бинарного выбора	Исходная выборка	11.00	42.86	18.3
	Выборка без аномальных наблюдений	4.00	30.77	12.5

Таким образом, алгоритм классификации на основе логит-модели бинарного выбора обладает лучшими прогностическими способностями по сравнению с методом дискриминантного анализа, о чем говорят меньшие значения вероятностей ошибок первого и второго рода до и после исключения аномальных наблюдений из выборки. Аномальные наблюдения были выявлены на этапе предварительного анализа данных с помощью методов визуализации данных и алгоритмов кластерного анализа. Рассматривалась также возможность исключения аномальных наблюдений с помощью двухэтапной процедуры, основанной на использовании теста Хампеля и межклассового расстояния Махалонобиса [3]. Исключение аномальных наблюдений из выборки уменьшает вероятности ошибочных классификаций, тем самым, увеличивая количество точных классификаций заемщиков. Во всех случаях вероятность ошибки первого рода намного меньше, чем вероятность ошибки второго рода, а это значит, что вероятность пропустить «проблемного» заемщика значительно ниже вероятности отнесения «непроблемного» заемщика к классу «проблемных».

Литература

1. *Гринь Н.В., Малюгин В.И.* Исследование точности методов классификации многомерных данных в задачах кредитного скоринга // Вестник ГрГУ. Сер.2. – 2008. – №1. – С.77–85.
2. *Малюгин В.И.* Оценка устойчивости банков на основе эконометрических моделей. Банковский Вестник, 2007, Февраль.
3. *Харин Ю. С.* Робастность в статистическом распознавании образов. Минск, Университетское, 1992.

КРИПТОГРАФИЧЕСКИЕ СВОЙСТВА ГЕНЕРАТОРА МАКЛАРЕНА – МАРСАЛЬИ

И. Б. Бережной

Для быстрой и надежной защиты больших объемов данных используются поточные криптосистемы, основным элементом которых